

PATENT APPLICATION
ATTORNEY DOCKET NO. ASC-060

GATE MATERIAL FOR SEMICONDUCTOR DEVICE
FABRICATION

Related Applications

This application claims the benefit of U.S. Provisional Application 60/420,227 filed October 22, 2002, the entire disclosure of which is hereby incorporated by reference.

Field of the Invention

This application relates generally to semiconductor devices and particularly to semiconductor structures made on semiconductor substrates with strained layers.

Background

Formation of metal-oxide-semiconductor field-effect transistors (MOSFETs) requires the introduction of dopants into, e.g., a silicon (Si) substrate to define source and drain regions. Dopants are also introduced into a gate material, such as polycrystalline silicon (polysilicon), to achieve a desired conductivity. Dopants disposed in source, drain, and gate (S/D/G) regions are activated by a heat treatment to provide the needed electrical characteristics. For good-quality n-type MOS (NMOS) devices, dopant activation is typically performed at a high temperature, e.g., at least 1000 °C for 5 seconds, to avoid polysilicon depletion effects. A gate in which the polysilicon is depleted has a non-uniform distribution of dopants, with a relatively low concentration of dopants near an interface with a gate dielectric. This depletion region can result in reduced gate capacitance during device operation, resulting in a lower transistor drive current.

Activation of dopants in regions defined on substrates with strained layers, such as strained Si, presents a challenge. Strained Si substrates include a thin strained Si layer having a thickness of, e.g., 40 - 400 Å. The strained Si layer is disposed over a second material, e.g., a relaxed SiGe layer. A compressively strained SiGe layer may be disposed above or below the strained Si layer. These layer structures may make it

difficult to maintain shallow source/drain junctions in, for example, complementary MOS (CMOS) devices, especially when the strained Si/SiGe substrate is subjected to high temperatures. This difficulty arises from the different diffusion rates of dopants in SiGe and in Si. For example, arsenic (As) may diffuse much more rapidly in SiGe than in Si at temperatures significantly above 900 °C and/or times significantly above 30 seconds.

This rapid diffusion leads to deeper source/drain junctions in NMOS transistors fabricated on strained-Si/SiGe substrates and/or excessive lateral diffusion of dopants beneath the gate, i.e., into the channel region. Because of the diffusion of As into a channel of the NMOS transistor, the transistor has a high off current (I_{off}) and it becomes more difficult to turn off.

In alternative structures, the second material over which the strained layer is disposed may be e.g., a bulk semiconductor substrate, or an insulating material. Here, too, it may be difficult to maintain shallow source/drain junctions or prevent excessive lateral diffusion of dopants, especially when the strained structures are subjected to high temperatures. This difficulty arises from the different diffusion rates of dopants in strained layers in comparison to bulk, non-strained materials. For example, boron diffuses faster through strained Si than through bulk Si.

A possible solution is to perform the S/D/G dopant activation at a restricted time and temperature (e.g., 900 °C for 30 sec). However, these restricted parameters may lead to unacceptable polysilicon depletion effects.

Summary

A semiconductor layer is pre-doped and a dopant distribution anneal is performed prior to gate definition. Subsequently formed shallow sources and drains, therefore, are not affected by the gate annealing step.

In an aspect, the invention features a method for forming a structure, the method including forming a layer over a substrate, the layer having a depletion region with a thickness less than approximately 20 angstroms. A portion of the layer is removed to define a gate of a transistor, the gate defining a channel length. A plurality of dopants are introduced into the substrate proximate the gate to define a source and a drain, and the substrate is heated to a temperature to activate the plurality of dopants. The temperature

is sufficiently low to prevent at least a portion of the plurality of dopants from diffusing enough to induce a high off current.

One or more of the following features may be included. The substrate may include an insulating layer. A strained layer may be disposed over the insulating layer.

5 The substrate may include a strained layer. The strained layer may be tensilely strained or compressively strained. The substrate may include a relaxed layer. The substrate may include germanium. The depletion region thickness may be less than 10 angstroms.

The induced off current may be less than 10^{-6} Amperes per micrometer, and preferably may be less than 10^{-9} Amperes per micrometer.

10 After the plurality of dopants are introduced, a portion of the plurality of dopants disposed in a region of the source may define a source extent proximate the channel, and after heating the substrate, the source extent may diffuse under the gate a distance extending less than 12.5% of the channel length. A concentration of the portion of dopants at the source extent may be at least approximately 10^{18} atoms/cubic centimeter.

15 After the plurality of dopants are introduced, a portion of the plurality of dopants disposed in a region of the drain may define a drain extent proximate the channel, and after heating the substrate, the drain extent may diffuse under the gate a distance extending less than 12.5% of the channel length. A concentration of the portion of dopants at the drain extent may be at least approximately 10^{18} atoms/cubic centimeter.

20 The layer may include a semiconductor and the step of forming the layer may include introducing a plurality of gate dopants into the layer, and heating the layer to a first temperature to alter a distribution of the gate dopants in the layer. The semiconductor may include silicon and/or germanium.

25 The layer may include a metallic element, such as at least one of molybdenum, titanium, tantalum, tungsten, iridium, nickel, cobalt, and platinum.

In another aspect, the invention features a method for forming a structure, the method including introducing a first plurality of dopants into a gate electrode layer disposed over a substrate. The gate electrode layer is heated to a first temperature to alter a distribution of the first plurality of dopants in the gate electrode layer. A portion of the
30 gate electrode layer is removed to define a gate of a transistor. A second plurality of dopants is introduced into the substrate proximate the gate to define a source and a drain.

The substrate is heated to a second temperature to activate the second plurality of dopants, with second temperature being less than the first temperature.

One or more of the following features may be included. The substrate may include an insulating layer. The substrate may include a strained layer disposed over the
5 insulating layer.

The substrate may include a strained layer. The strained layer may be tensilely strained or compressively strained. The substrate may include a relaxed layer. The substrate may include germanium.

The first temperature may be greater than 1000 °C. The second temperature may
10 be less than 1000 °C.

The gate electrode layer may include a semiconductor layer, such as silicon and/or germanium.

The first plurality and the second plurality of dopants may include n-type dopants and/or or p-type dopants.

15 In another aspect, the invention features a method for forming a structure, the method including introducing a first plurality of dopants into a gate electrode layer disposed over a substrate. The semiconductor layer is heated for a first time period to alter a distribution of the first plurality of dopants in the gate electrode layer. A portion of the gate electrode layer is removed to define a gate of a transistor. A second plurality
20 of dopants is introduced into the substrate proximate the gate to define a source and a drain. The substrate is heated for a second time period to activate the second plurality of dopants, with the second time period having a shorter duration than a duration of the first time period.

One or more of the following features may be included. The substrate may
25 include an insulating layer. The substrate may include a strained layer disposed over the insulating layer. The substrate may include a strained layer. The strained layer may be tensilely strained or compressively strained. The substrate may include a relaxed layer. The substrate may include at least one of silicon and germanium.

The first time period may be greater than 5 seconds. In some embodiments, the
30 first time period may be greater than 30 seconds.

The gate electrode layer may include a semiconductor layer. The semiconductor layer may include silicon and/or germanium.

The first and the second plurality of dopants may include n-type dopants and/or p-type dopants.

5 In another aspect, the invention features a structure including a strained layer disposed over a substrate. A first transistor includes a first source and a first drain, with at least a portion of the first source and the first drain disposed in a first portion of the strained layer. The first gate is disposed above the strained layer and between the source and drain regions, the first gate including a first metal. A first gate dielectric layer is
10 disposed between the first gate and the strained layer.

One or more of the following features may be included. The substrate may include dielectric material and the strained layer may be disposed in contact with the dielectric material. The first metal may be selected from the group consisting of titanium, tungsten, molybdenum, tantalum, nickel, cobalt, and platinum. The strained layer may
15 include silicon and/or germanium. The gate may include a metal-semiconductor alloy. In some embodiments, the gate may include only metal silicide.

A channel may be disposed under the gate. The source may include a source extent proximate the channel, the source extent extending under the gate a distance less than 12.5% of a channel length. A concentration of dopants in the source extent may be
20 at least approximately 10^{18} atoms/cubic centimeter. The drain may include a drain extent proximate the channel, the drain extent extending under the gate a distance less than 12.5% of a channel length. A concentration of dopants in the drain extent may be at least approximately 10^{18} atoms/cubic centimeter.

The structure may have a second transistor that includes a second source and a
25 second drain, with at least a portion of the first source and the first drain disposed in a second portion of the strained layer. A second gate may be disposed above the strained layer and between the second source and second drain regions, the second gate including a second metal. A second gate dielectric layer may be disposed between the second gate and the strained layer. The first transistor may be an n-type metal-oxide semiconductor
30 field-effect transistor, the first source and the first drain may include n-type dopants, the

second transistor may be a p-type metal-oxide-semiconductor field-effect transistor, and the second source and second drain may include p-type dopants.

The first gate may have a first workfunction, the second gate may have a second workfunction, and the first workfunction may be substantially equal to or substantially
5 different from the second workfunction.

Brief Description of the Drawings

Figures 1 - 4 illustrate several substrates amenable for use fabrication of semiconductor structures;

Figures 5 - 8A are a series of schematic cross-sectional views of a semiconductor
10 substrate illustrating a process for fabricating a semiconductor structure on the substrate;

Figure 8B graphically depicts a distribution of dopants in the semiconductor structure illustrated in Figure 8A;

Figure 9 is a schematic cross-sectional view of a semiconductor structure fabricated on the substrate; and

15 Figure 10 is a schematic cross-sectional view of a semiconductor structure fabricated on another substrate.

Like-referenced features represent common features in corresponding drawings.

Detailed Description

20 Referring to Figure 1, which illustrates an epitaxial wafer 100 amenable to use with the present invention, several layers collectively indicated at 101, including a strained layer 102 and a relaxed layer 104, are disposed over a substrate 106. The ensuing discussion focuses on a strained layer 102 that is tensilely strained, but it is understood that strained layer 102 may be tensilely or compressively strained. Strained
25 layer 102 has a lattice constant other than the equilibrium lattice constant of the material from which it is formed, and it may be tensilely or compressively strained; relaxed layer 104 has a lattice constant equal to the equilibrium lattice constant of the material from which it is formed. Tensilely strained layer 102 shares an interface 108 with relaxed layer 104.

30 Substrate 106 and relaxed layer 104 may be formed from various materials systems, including various combinations of group II, group III, group IV, group V, and

group VI elements. For example, each of substrate 106 and relaxed layer 104 may include a III-V compound. Substrate 106 may include gallium arsenide (GaAs), and relaxed layer 104 may include indium gallium arsenide (InGaAs) or aluminum gallium arsenide (AlGaAs). These examples are merely illustrative, and many other material systems are suitable.

In an embodiment, relaxed layer 104 may include $\text{Si}_{1-x}\text{Ge}_x$ with a uniform composition, containing, for example, Ge in the range $0.1 \leq x \leq 0.9$ and having a thickness T_1 of, e.g., 0.2 - 2 μm . In an embodiment, T_1 is 1.5 μm .

Strained layer 102 may include a semiconductor such as at least one of a group II, a group III, a group IV, a group V, and a group VI element. Strained semiconductor layer 102 may include, for example, Si, Ge, SiGe, GaAs, indium phosphide (InP), and/or zinc selenide (ZnSe). In some embodiments, strained semiconductor layer 102 may include approximately 100% Ge, and may be compressively strained. A strained semiconductor layer 102 comprising 100% Ge may be formed over, e.g., relaxed layer 104 containing uniform $\text{Si}_{1-x}\text{Ge}_x$ having a Ge content of, for example, 50 - 90 % (i.e., $x = 0.5 - 0.9$), preferably 70% (i.e., $x = 0.7$).

In an embodiment, tensilely strained layer 102 is formed of silicon. Tensilely strained layer 102 has a thickness T_2 of, for example, 50 - 1000 Å. In an embodiment, thickness T_2 is less than 200 Å.

Relaxed layer 104 and strained layer 102 may be formed by epitaxy, such as by atmospheric-pressure CVD (APCVD), low- (or reduced-) pressure CVD (LPCVD), ultra-high-vacuum CVD (UHVCVD), by molecular beam epitaxy (MBE), or by atomic layer deposition (ALD). Strained layer 102 containing Si may be formed by CVD with precursors such as silane, disilane, or trisilane. Strained layer 102 containing Ge may be formed by CVD with precursors such as germane or digermane. The epitaxial growth system may be a single-wafer or multiple-wafer batch reactor. The growth system may also utilize a low-energy plasma to enhance layer growth kinetics.

In an embodiment in which strained layer 102 contains substantially 100% Si, strained layer 102 may be formed in a dedicated chamber of a deposition tool that is not exposed to Ge source gases, thereby avoiding cross-contamination and improving the quality of interface 108 between strained layer 102 and relaxed layer 104. Furthermore,

strained layer 102 may be formed from an isotopically pure silicon precursor(s).

Isotopically pure Si has better thermal conductivity than conventional Si. Higher thermal conductivity may help dissipate heat from devices subsequently formed on strained layer 102, thereby maintaining the enhanced carrier mobilities provided by strained layer 102.

5 In some embodiments, relaxed layer 104 and/or strained layer 102 may be planarized by, e.g., CMP, to improve the quality of subsequent wafer bonding. Strained layer 102 may have a low surface roughness, e.g., less than 0.5 nanometer (nm) root mean square (RMS).

10 Referring to Figure 2, an alternative epitaxial wafer 100 amenable for use with the present invention may include layers in addition to those indicated in Figure 1. For example, a substrate 200 formed from a semiconductor, such as silicon, may have several layers collectively indicated at 202 formed upon it. Layers 202 may be grown, for example, by APCVD, LPCVD, or UHVCVD.

15 Layers 202 include a graded layer 204 disposed over substrate 200. Graded layer 204 may include Si and Ge with a grading rate of, for example, 10% Ge per μm of thickness, and a thickness T_3 of, for example, 2 - 9 μm . Graded layer 204 may be grown, for example, at 600 - 1200 $^{\circ}\text{C}$. See, e.g., U.S. Patent No. 5,221,413, incorporated herein by reference in its entirety. Relaxed layer 104 is disposed over graded layer 204. A virtual substrate 206 includes relaxed layer 104 and graded layer 204.

20 A compressively strained layer 208 including a semiconductor material is disposed over relaxed layer 104. In an embodiment, compressively strained layer 208 includes group IV elements, such as $\text{Si}_{1-y}\text{Ge}_y$, with a Ge content (y) higher than the Ge content (x) of relaxed $\text{Si}_{1-x}\text{Ge}_x$ layer 104. Compressively strained layer 208 contains, for example, Ge in the range $0.25 \leq y \leq 1$ and has a thickness T_4 of, e.g., 10 - 500 angstroms (Å).
25 (Å). In some embodiments, compressively strained layer 208 has a thickness T_4 of less than 500 Å. In certain embodiments, T_4 is less than 200 Å.

Tensilely strained layer 102 is disposed over compressively strained layer 208, sharing an interface 210 with compressively strained layer 208. In some embodiments, compressively strained layer 208 may be disposed not under, but over tensilely strained
30 layer 102.

Substrate 200 with layers 202 typically has a threading dislocation density of 10^4 - $10^5/\text{cm}^2$.

Referring to Figure 3, yet another alternative epitaxial wafer amenable for use with the present invention is a strained-semiconductor-on-semiconductor SSOS substrate 300, having a strained layer 102 disposed in contact with a crystalline semiconductor handle wafer 310. Handle wafer 310 may include a bulk semiconductor material, such as silicon. The strain of strained layer 102 is not induced by underlying handle wafer 310, and is independent of any lattice mismatch between strained layer 102 and handle wafer 310. In an embodiment, strained layer 102 and handle wafer 310 include the same semiconductor material, e.g., silicon. Handle wafer 310 may have a lattice constant equal to a lattice constant of strained layer 102 in the absence of strain. Strained layer 102 may have a strain greater than approximately 10^{-3} . Strained layer 102 may have been formed by epitaxy, and may have a thickness T_2 ranging from approximately 20 Å to approximately 1000 Å, with a thickness uniformity of better than approximately $\pm 10\%$. In an embodiment, strained layer 102 may have a thickness uniformity of better than approximately $\pm 5\%$. Strained layer 102 may have a surface roughness of less than 20 Å.

The SSOS substrate 300 may be formed, as described in U.S. Serial Nos. 10/456,708, 10/456,103, 10/264,935, and 10/629,498, the entire disclosures of each of the four applications being incorporated herein by reference. The SSOS substrate formation process may include the formation of strained layer 102 over substrate 106 as described above with reference to Figure 1. A cleave plane may be defined in, e.g., relaxed layer 104. Strained layer 102 may be bonded to the handle wafer 310, and a split may be induced at the cleave plane. Portions of the relaxed layer 104 remaining on strained layer 102 may be removed by, e.g., oxidation and/or wet etching.

Yet another epitaxial wafer amenable for use with the present invention is a strained-semiconductor-on-insulator (SSOI) wafer 400. Referring to Figure 4, a SSOI wafer 400 has strained layer 102 disposed over an insulator, such as a dielectric layer 410 formed on a semiconductor substrate 420. SSOI substrate 400 may be formed by methods analogous to the methods described above in the formation of SSOS substrate 300. Dielectric layer 410 may include, for example, SiO_2 . In an embodiment, dielectric layer 410 includes a material having a melting point (T_m) higher than a T_m of pure SiO_2 ,

i.e., higher than 1700 °C. Examples of such materials are silicon nitride (Si_3N_4), aluminum oxide, magnesium oxide, etc. In another embodiment, dielectric layer 410 includes a high-k material with a dielectric constant higher than that of SiO_2 , such as aluminum oxide (Al_2O_3), hafnium oxide (HfO_2) or hafnium silicate (HfSiON or HfSiO_4).

5 Semiconductor substrate 420 includes a semiconductor material such as, for example, Si, Ge, or SiGe. Strained layer 102 has a thickness T_4 selected from a range of, for example, 50 - 1000 Å, with a thickness uniformity of better than approximately $\pm 5\%$ and a surface roughness of less than approximately 20 Å. Dielectric layer 410 has a thickness T_5

- 10 102 includes approximately 100% Si or 100% Ge having one or more of the following material characteristics: misfit dislocation density of, e.g., 0 - 10^5 cm/cm²; a threading dislocation density of about 10^1 - 10^7 dislocations/cm²; a surface roughness of approximately 0.01 - 1 nm RMS; and a thickness uniformity across SSOI substrate 400 of better than approximately $\pm 10\%$ of a mean desired thickness; and a thickness T_4 of less
15 than approximately 200 Å. In an embodiment, SSOI substrate 400 has a thickness uniformity of better than approximately $\pm 5\%$ of a mean desired thickness.

In an embodiment, dielectric layer 410 has a T_m greater than that of SiO_2 . During subsequent processing, e.g., MOSFET formation, SSOI substrate 400 may be subjected to high temperatures, i.e., up to 1100 °C. High temperatures may result in the relaxation of
20 strained layer 102 at an interface 430 between strained layer 102 and dielectric layer 410. The use of dielectric layer with a T_m greater than 1700 °C may help keep strained layer 102 from relaxing at the interface 430 between strained layer 102 and dielectric layer 410 when SSOI substrate is subjected to high temperatures.

In an embodiment, the misfit dislocation density of strained layer 102 may be
25 lower than its initial dislocation density. The initial dislocation density may be lowered by, for example, performing an etch of a top surface 440 of strained layer 102. This etch may be a wet etch, such as a standard microelectronics clean step such as an RCA SC1, i.e., hydrogen peroxide, ammonium hydroxide, and water ($\text{H}_2\text{O}_2 + \text{NH}_4\text{OH} + \text{H}_2\text{O}$), which at, e.g., 80 °C may remove silicon.

30 In an embodiment, substrate 210 with layers 202 is processed through various CMOS front-end steps such as well definition and isolation formation (not shown).

Referring to Figure 5, a gate dielectric layer 500 is formed on a top surface 510 of strained layer 102. Gate dielectric layer 500 is, for example, a thermally grown gate oxide such as silicon dioxide (SiO_2). Alternatively, gate dielectric layer 500 may include a high-k material with a dielectric constant higher than that of SiO_2 , such as aluminum oxide (Al_2O_3), hafnium oxide (HfO_2) or hafnium silicate (HfSiON or HfSiO_4). In some embodiments, gate dielectric layer 500 may be a stacked structure, e.g., a thin SiO_2 layer capped with a high-k material. A gate electrode layer 520 is formed over gate dielectric layer 500. Gate electrode layer 520 may include, for example, polysilicon, amorphous silicon, Ge, or SiGe gate material.

Referring to Figure 6, an implantation mask 600 is formed over gate electrode layer 520. Implantation mask 600 may be made of a masking material such as photoresist. Implantation mask 600 defines an opening 610, with opening 610 exposing a portion 620 of gate electrode layer 520 (defined for purposes of illustration by the dashed lines). Gate electrode layer portion 620 is disposed over a portion of region 630 of substrate 200 and layers 202 in which NMOS devices will be formed. Implantation mask 600 protects portions of the top surface 640 of gate electrode layer 520 disposed over regions of substrate 200 and layers 202 in which NMOS devices will not be formed. In the illustrated embodiment, implantation mask exposes only an area 620 in which an NMOS gate will be defined (see below). In some other embodiments, implantation mask 600 exposes entire regions of gate electrode layer 520 disposed over regions of substrate 200 and layers 202 in which NMOS devices will be formed, including regions in which n-type sources and drains will be formed (see below).

Subsequent to the formation of implantation mask 600, a plurality of n-type dopants 650 are introduced into gate electrode layer portion 620 through opening 610. N-type dopants 650 may be, for example, As or phosphorus (P) ions introduced by ion implantation. After the implantation of n-type dopants, implantation mask 600 is removed by a stripping process such as a dry strip in an oxygen plasma. A diffusion anneal is performed to diffuse n-type dopants 650 uniformly in a vertical direction throughout portion 620 of gate electrode layer 520. This diffusion anneal is performed at a relatively high temperature, e.g., over 1000°C , such as 1025°C , for a sufficiently long time to uniformly diffuse dopants 650, e.g., 5 seconds or more. The diffusion anneal

results in the formation of a depletion region 660 in portion 620 of gate electrode layer 520 having a thickness T_6 of, e.g., less than 20 angstroms, preferably less than 10 angstroms.

Referring to Figure 7, as well as to Figure 6, a gate 700 formed from gate electrode layer 520 is defined as follows. A gate photoresist mask (not shown) is deposited and patterned to protect at least part of portion 620 of gate electrode layer 520. Regions of gate electrode layer 520, as well as regions of portion 620, exposed by the gate photoresist mask are removed by a removal process such as reactive ion etching (RIE). Subsequently, portions of dielectric layer 500 exposed by the RIE of portions of gate electrode layer 520 are also removed by a removal step, such as RIE with an etch chemistry selective to the material comprising strained layer 102, such as Si. Removal of portions of dielectric layer 500 exposes top surface 510 of strained layer 102, and defines a gate dielectric 710 disposed under gate 700. The gate photoresist mask is removed by, for example, a stripping process such as a dry strip in an oxygen plasma. Gate 700 includes a uniform distribution of n-type dopants, and defines an initial channel length L_1 .

Referring to Figure 8A, a shallow implantation of n-type dopants, such as As, is performed to define a source extension 800 and a drain extension 810 in strained layer 102. A first sidewall spacer 820 and a second sidewall spacer 830 are defined proximate gate 700. First and second sidewall spacers 820, 830 are formed from a dielectric, such as silicon dioxide or silicon nitride. A source 840 and a drain 850 may be defined in portions of strained layer 102, compressively strained layer 208, and relaxed layer 104, proximate first and second sidewall spacers 820, 830. In some embodiments, source 840 and drain 850 may be defined in strained layer 102. Source 840 and drain 850 are defined by the introduction of a plurality of dopants, such as by an implantation of n-type dopants, e.g., As, into layers 202 disposed over substrate 200. These dopants are substantially prevented from reaching regions of compressively strained layer 208 and strained layer 102 disposed below gate dielectric 710 by the presence of sidewall spacers 820, 830.

After the introduction of dopants to define source 840, drain 850, source extension 800, and drain extension 810, an activation anneal is performed to activate

these dopants. The activation anneal is performed at a relatively low temperature, e.g., less than 1000 °C. For example, an activation anneal may be done at 900 °C for 30 seconds. Alternatively, the activation anneal may be done for a very short duration at a higher temperature, e.g., 1 second at 1100 °C. In an alternative embodiment, an

5 activation anneal of extremely short duration (e.g., less than 1 second) may be performed by techniques such as flash lamp annealing or laser annealing at temperatures between 900 °C and 1350 °C. This temperature and time are sufficient to activate the dopants in the source 840 and drain 850, without inducing excessive diffusion of n-type dopants into a channel 860 under gate 700. As a result of this procedure, good dopant activation is

10 achieved and polysilicon depletion avoided due to the high-temperature diffusion anneal. At the same time, dopants in the vicinity of strained layer 102 and/or compressively strained layer 208 do not experience high temperatures for long durations (high thermal budgets) and, hence, do not significantly invade these layers beyond the boundaries of source 840, drain 850, source extension 800, and drain extension 810. Dopants do not

15 diffuse into channel 860 enough to induce a high off current. The off current may be less than 10^{-6} Amperes per micrometer. In some embodiments, the off current is less than 10^{-9} Amperes per micrometer.

Referring to Figure 8B as well as to Figure 8A, a concentration of dopants in layers 202 may be graphically depicted with a graph 865, with an x-axis representing

20 positions within strained layer 102 and a y-axis representing a logarithm of dopant concentration. A concentration [n] 870 of dopants in source 840 and source extension 800 and a concentration [n] 875 of dopants in drain 850 and drain extension 810 may have a maximum level of approximately 10^{21} atoms/cubic centimeter at a level 880 disposed below a surface of strained layer 102. Dopants disposed in an outer region of

25 source 840 may define a source extent 890, and dopants disposed in an outer region of drain 850 may define a drain extent 895. The concentration of dopants at source extent 76 and drain extent 78 may be approximately 10^{18} atoms/cubic centimeter. After heating of substrate 200, portions of source extent 890 and drain extent 895 disposed proximate channel 860 may diffuse a distance extending less than 12.5% of gate length L_1 , thereby

30 decreasing channel length L_1 by no more than 25%. The abruptness of the dopant concentration in the source and drain region may also be greater than 2 nm per decade

(i.e., per order of magnitude in concentration). In some embodiments, this abruptness may be better than 4 nm/decade.

In an alternative embodiment, PMOS devices are formed with pre-doped gates. Here, the semiconductor material from which the PMOS gate will be defined is doped
5 with p-type dopants (e.g. boron or indium) prior to PMOS gate definition.

In some embodiments, source and drain extensions may extend into an underlying layer that may include an element other than Si, such as Ge.

In an alternative embodiment, no mask, e.g., no implantation mask 600, is formed before gate electrode layer 520 is implanted with n-type dopants. In some applications,
10 however, implantation of n-type dopants into gate electrode layer 520 material which will be used for PMOS gates (or vice versa) may adversely affect threshold voltages.

In some embodiments, gate 700 is formed from a conductive material that does not require doping, such as a metal. Gate 700 can be formed from metals such as titanium (Ti), tungsten (W), molybdenum (Mo), or tantalum (Ta), as well as other
15 materials, e.g., titanium nitride (TiN), titanium silicon nitride (TiSiN), tungsten nitride (WN), tantalum nitride (TaN), tantalum silicide (TaSi), iridium (Ir), iridium oxide (IrO₂), etc., that provide an appropriate workfunction, i.e., a workfunction of approximately 4 - 5.5 electron volts (eV), without doping. Metal gates may have a depletion region of 20 angstroms or less, preferably less than 10 angstroms.

Referring to Figure 9, a first transistor 910 and a second transistor 920 may be
20 formed over strained layer 102. At least a portion 922 of first source 840 and at least a portion 924 of first drain 850 may be disposed in a first portion 930 of the strained layer 102. First source 840 and first drain 924 may extend into compressively strained layer 208 and relaxed layer 104. First gate 700 may be disposed above strained layer 102 and
25 between first source 840 and first drain 850. First gate 700 may include a metal, such as titanium, tungsten, molybdenum, tantalum, nickel, cobalt, or platinum. In some embodiments, gate 700 may contain a metal-semiconductor alloy, such as metal silicide, metal germanocide, or metal germanosilicide. In some embodiments, the gate 700 may include only a metal-semiconductor alloy. Channel 860 may be disposed under gate 700.
30 Source 840 may include source extent 890 and drain 850 may include a drain extent 895. Each or both of the drain source extent 890 and drain extent 895 may extend under gate

700 a distance less than 12.5% of channel length L_1 (see Figure 8A). A concentration of dopants in the source extent 890 and/or the drain extent 895 may be at least 10^{18} atoms/cubic centimeter.

The second transistor 920 may include a second source 940 and a second drain 950 disposed in a second portion 960 of the strained layer 102. A second gate 965 may be disposed above the strained layer 102 and between the second source 940 and second drain 950. The second gate 965 may include a second metal, such as titanium, tungsten, molybdenum, tantalum, nickel, cobalt, or platinum. In some embodiments, gate 965 may contain a metal-semiconductor alloy. In some embodiments, the gate 965 may include only a metal-semiconductor alloy. A second gate dielectric layer 970 may be disposed between the second gate 965 and the strained layer 102. The first transistor 910 may be an n-type metal-oxide semiconductor field-effect transistor (n-MOSFET), the first source 840 and the first drain 850 may include n-type dopants. The second transistor 920 may be a p-type metal-oxide-semiconductor field-effect transistor (p-MOSFET), and the second source 940 and second drain 950 may include p-type dopants. CMOS device 900, therefore, includes both n-MOSFET 910 and p-MOSFET 920.

In some embodiments, gates 700 and 965 may be formed from semiconductor layers or from metal-semiconductor alloys, such as silicides.

In a CMOS device, a single gate having a mid-band gap workfunction (approximately 4.4 - 4.6 eV) may be used for both NMOS and PMOS devices such as, for example, fully depleted semiconductor-on-insulator devices built on SSOI substrates. Alternatively, two different materials having workfunctions closer to the respective band edges, e.g., approximately 0.2 - 0.4 eV below the conduction band edge (~ 4 eV) or approximately 0 - 0.2 eV above the valence band edge (~ 5 eV), may be used for NMOS and PMOS devices, respectively, formed with strained semiconductors such as strained silicon. By using a gate material that provides an appropriate workfunction without doping, gate depletion effects are avoided because dopants are unnecessary. Further, adverse short channel effects due to dopant diffusion from source and drain extensions 800, 810 are also avoided by the elimination of high thermal budget activation steps.

In some embodiments, gate electrode layer 520 may be formed from a gate semiconductor material such as polycrystalline Si, Ge, or SiGe that is reacted with a

subsequently deposited metal, e.g., nickel, cobalt, titanium, or platinum, either before or after the definition of gate 700. The gate semiconductor material may be deposited as a layer by CVD and may have a thickness of approximately 500 - 2000 Å, e.g., 1000 Å. The subsequently deposited metal may be deposited by, e.g., sputter deposition, and may have a thickness of, for example, 2 - 15 nm. The gate semiconductor material and the metal may be reacted in a reaction process such as a silicidation process that includes, e.g., rapid thermal processing at, for example, 10 - 120 seconds at 400 - 850 °C. The reaction process can also include a second rapid thermal processing step after a wet chemical strip that removes any unreacted metal from the structure. In these embodiments, the reaction conditions and the thicknesses of the gate semiconductor material and the metal are selected such that the gate semiconductor material and the metal substantially completely react with each other to form a metal-semiconductor alloy, such as a metal silicide. Gate 700, thus, substantially comprises a silicide material such as nickel silicide, cobalt silicide, titanium silicide, or platinum silicide, or a germanocide material such as nickel germanocide, cobalt germanocide, titanium germanocide, or platinum germanocide. Gate electrode layer 520 may be doped by the introduction of a plurality of n-type or p-type dopants prior to the reaction process, e.g., the silicidation reaction. Such doping may alter the post-reaction process gate workfunction, facilitating the fabrication of devices with a desired threshold voltage. The reaction process may be performed before or after the definition of gate 700. Because the full reaction of the semiconductor gate electrode layer 520 (and hence gate 700) results in gate 700 being a metal gate, polysilicon depletion effects are eliminated.

Referring to Figure 10 as well as to Figures 4 and 9, transistor 910 may be formed over SSOI substrate 410, in which strained layer 102 is disposed in contact with dielectric layer 410. In this embodiment, source 840 and drain 850 are disposed entirely within strained layer 102.

The methods and structures described above with reference to Figures 5 - 10 may be formed on other epitaxial wafers, such as the wafers illustrated in Figures 1 and 3.

The invention may be embodied in other specific forms without departing from the spirit or essential characteristics thereof. The foregoing embodiments are therefore to be considered in all respects illustrative rather than limiting on the invention described

herein. Scope of invention is thus indicated by the appended claims rather than by the foregoing description, and all changes which come within the meaning and range of equivalency of the claims are intended to be embraced therein.

What is claimed is: